UNIVERSITY OF FORWARD THINKING WESTMINSTER

Submission to House of Lords Select Committee on Artificial Intelligence

Bunz, M.

Submission to House of Lords Select Committee on Artificial Intelligence

Dr. Mercedes Bunz, University of Westminster

My background and expertise

My research explores the effects of algorithms on knowledge. I studied recent AI developments for the book 'The Internet of Things' (Polity, Nov 2017, co-authored with Graham Meikle) and analysed this topic in a previous book 'The Silent Revolution: How Digitalization Transforms Knowledge, Work, Journalism and Politics without Making Too Much Noise' (Palgrave Macmillan 2014). As the former technology reporter of *The Guardian*, I also have significant knowledge of start-up culture and have just become member of the Internet of Things Working Group organized by the *British Standards Institute*.

Summary of evidence

AI creates a new data paradigm. To advance this kind of AI, the role of government can make quite a difference. By processing data AI programs *create knowledge* i.e. they take over knowledge tasks and even make decisions. To develop and train this kind of AI, large amounts of data are needed in sufficient size and quality. For the creation of this data, the decisions of government are vital. Policy initiatives should: A) ensure that the UK has a strategy for the creation of big datasets in high quality; B) identify sectors in which the creation of those datasets is especially desirable; C) create incentives for businesses to open and share their datasets; D) foster data sovereignty to minimize UK's algorithmic dependence on U.S. products.

1. Datasets are crucial to train AI. A new programming technique known as '*neural networks'* created the current breakthrough in AI technology. Instead of programming rules that should be applied, a neural network *infers rules of categorization* by analyzing correctly labelled data records in the thousands. To train neural networks in this mode of 'deep learning', computer scientists depend on large datasets. AI's ability to recognize objects in images and videos, for example, has evolved with ever larger image-language datasets on which the objects or actions displayed are correctly named. In the recent past, the UK ran an image recognition challenge based on a dataset known as *PASCAL Voc* (2005 – 2012)[1], which allowed computers to train on about 20,000 annotated images. This was soon being

[1] The PASCAL Visual Object Classes Homepage http://host.robots.ox.ac.uk/pascal/VOC/ was organised by Mark Everingham (University of Leeds), Luc van Gool (ETHZ, Zurich), Chris Williams (University of Edinburgh), John Winn (Microsoft Research Cambridge) and Andrew Zisserman (University of Oxford).

surpassed by datasets created in the U.S. such as the *Flickr30k*[2] set with over 30,000 pictures focusing on people and everyday activities, and Microsoft's *MS COCO* consisting of 300,000 images offering multiple objects per image. Around 2009, the up-to-date largest dataset was created with the University of Stanford's *ImageNet*, which provided over one million images with annotations. *To train and test AI, datasets must be huge.*

2. AI creates a new data paradigm. As the report '*The big data dilemma'* [3] by the Science and Technology Committee has documented, the UK government is aware of the importance of Open Data and of opening government data. In times of AI, however, this data is not just useful for informing businesses and for allowing start-ups to create new services. Datasets are now also essential to train AI that then can create further knowledge. As AI needs to be trained, only areas for which datasets are available can advance. This is making *a data agenda* necessary. AI has created a new data paradigm – from *public infrastructure* to the *creation of knowledge*.

3. The creation of datasets is expensive. Financially, they depend on research funding or have mostly been created by large corporations such as Google, Facebook etc. Data is available everywhere, but datasets to train AI need to be 'cleaned', i.e. labelled and corrected. In the past, AI research projects have turned to Amazon's micro-task marketplace Mechanical Turk.[4] *ImageNet* employed at its peak 48,940 people in 167 countries that sorted and labelled nearly a billion images downloaded from the internet. According to its former director, Stanford Professor Fei-Fei Li, it was at one point one of Mechanical Turk's biggest employers. Its usage in the *ImageNet Challenge* led to massive breakthroughs in AI-driven image recognition. This is a clear indication that data and the creation of datasets and even more their constant update is as laborious and costly as it is central to developing and training AI. *The creation of datasets, laborious and expensive, is essential to advance AI.*

4. A lot of data is siloed and proprietary. Citizens create data every day in public and commercial environments. However, this data is often out of reach for start-ups and programmers. While programmers do know that a needed dataset exists in another company, it is proprietary and cannot be used to train or update their AI. As the Royal Academy of Engineering stated in its 2015 *Connecting Data* report quoted by the Select Committee: 'Much potentially valuable data remains locked away in corporate silos or within

[2] Flickr30K is hosted by the University of Illinois https://illinois.edu/fb/sec/229675.
[3] Science and Technology Committee, House of Commons, *The big data dilemma: Fourth Report of Session 2015–16*, 10 February 2016.
[4] As described in: Karpathy, A. & Fei-Fei, L. (2015) 'Deep visual-semantic alignments for generating image descriptions', *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3128–37.

sectors'.[5] New developments such as the internet of things will mean that more and more data of citizens will be collected in that proprietary manner. 'Data portability' – allowing individuals to re-use their personal data – and voluntary programs such as UK's 'midata' initiative support this. But individual portability will not be sufficient to collect a dataset that allows the creation of knowledge and businesses. The UK needs a strategy to actively create big data, especially in areas of government interest such as healthcare, transport, science and education. Citizens creating data is of value. The data of 1.6 million patients of the Royal Free Hospital that was given to the company DeepMind has made headlines for breaching UK data law, while the immense value of that dataset has been overlooked. *The creation of open datasets should be actively pursued. Specific areas of interest should be identified in a data agenda.*

5. AI frameworks are often bought as a service from U.S. companies. As UK companies and research institutions experience difficulties to get their hands on big UK data, they are struggling to train their own AI-driven programs. This is especially the case in the field of natural language processing, in which U.S. companies got a head start with the collection of data. Many of UK's digital businesses working with conversational technology such as chatbots bought them from one of the five U.S. companies dominating this market: Google, IBM's Watson, Amazon's Lex, Microsoft, or Facebook.[6] Their chatbot frameworks provide natural language processing abilities that are then further specified by UK companies for specific task by adding on 'domain knowledge'. This can mean that UK services who use U.S. chatbots send their data back to U.S. servers including in sensitive areas such finance (UK banks using chatbots) or healthcare. *Data sovereignty should be a subject addressed in the data governance framework.*

6. AIs are prone to be biased. The Science and Technology Committee has noted this in their '*Robotics and artificial intelligence report'*[7] as being of ethical and legal concern. In face of the rapid applications of AI in a range of sensitive sectors such as news production (the BBC and the Press Association are currently working on AI projects) or healthcare (Deepmind, Babylon Health and others) a government strategy to address potential bias is needed. AI learns to categorize by looking for patterns, and can easily amplify existing biases resulting from biased training data or by an insufficient training of the AI. *Recommendations for the creation of datasets could help to minimise bias.*

---

[5] Science and Technology Committee, House of Commons, *The big data dilemma: Fourth Report of Session 2015–16*, February 2016, p. 25, par 53).
[6] Conversational technology is currently offered by Google, IBM's Watson, Amazon's Lex, Microsoft and Facebook.
[7] Science and Technology Committee, House of Commons, '*Robotics and artificial intelligence report: Fifth Report of Session 2016–17*, September 2016.

7. Ensuring privacy and consent is another important ethical concern noted in the Science and Technology Committee's report. New approaches such as 'Differential Privacy' could be explored and recommended if proven reliable. *Differential Privacy* adds mathematical noise to a small sample of the individual's usage pattern to obscure an individual's identity without statistically harming the general pattern. *Today, a large number of statistical analyses can already be done in a differentially private manner by adding little noise.*

8. By actively providing datasets and/or alternatively a strong framework for the creation and maintenance of datasets, the government could address the above discussed issues of privacy, bias and data sovereignty. Especially in sensitive areas (such as healthcare and banking) but also in areas of public interest (such as transport and journalism) a framework for *the creation and the maintenance of data can help to avoid problems of bias and privacy that would undermine public trust.*

9. Identifying areas of data interest. The government's policy paper '*7. Data - unlocking the power of data in the UK economy and improving public confidence in its use'*[8] is a step in the right direction. To push this agenda further and to ensure that the UK remains at the forefront of data innovation, the government needs to actively identify areas and ways in which *the sharing of datasets should be encouraged.* EU-funded accelerators such as *digitalhealth* currently assist private businesses by providing in-depth knowledge of the NHS. This should become a two-way street especially in areas of public interest, with businesses to be encouraged to share data and knowledge. *It is important to create incentives for businesses to open and share their datasets.*

10. AI is a programming paradigm that is performing knowledge tasks and therefore takes part in the creation of knowledge. A general condition for this new knowledge is the availability of data. In the interest of UK citizens and businesses*, the availability of data should be stimulated in some areas and ensured in others.*

Mercedes Bunz, 1 September 2017.

Contact details:
Email: m.bunz@westminster.ac.uk
Phone: 077 206 33193

[8] Department for Digital, Culture, Media & Sport, 7. Data - unlocking the power of data in the UK economy and improving public confidence in its use, March 2017.